DOCUMENT RESUME

ED 400 283 TM 025 551

AUTHOR Gershon, Richard C.
TITLE Dissecting Item Misfit on Vocabulary Items.
PUB DATE Apr 91
NOTE 14p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1991).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Ability; Difficulty Level; *Goodness of Fit; *Item Response Theory; Test Construction; *Test Items; *Vocabulary
IDENTIFIERS Calibration; Johnson O Connor Aptitude Tests; *Rasch Model

ABSTRACT

The Johnson O'Connor Research Foundation, which produces vocabulary instructional materials for test takers, is in the process of determining the difficulty values of nontechnical words in the English language. To this end, the Foundation writes test items for vocabulary words and tests them in schools. The items are then calibrated using the Rasch model. This procedure results in a significant number of items being labeled as misfitting and being rejected from the item bank. A mislead analysis technique was created to try to uncover the sources of problems in items with poor fit statistics. The dataset used contained test results for over 3,500 items, each of which was administered to 400 to 600 persons, for a total of approximately 23,000 persons. General mislead curves were compared to the actual performance for items previously labeled as misfitting, and a mislead characteristic curve was established. A mislead table was constructed for each item. The mislead was considered to be significantly flawed for a given ability group when the observed performance differed from the means by more than two standard deviations. Each cell in the mislead table was evaluated in this way, giving item writers a way to observe which item choices are not functioning as expected. Five appendixes give examples of the mislead profiles for specific words. (Contains one figure and one table.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 400 283

Dissecting Item Misfit on Vocabulary Items

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
BETTY BERGSTROM
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Richard C. Gershon

Computer Adaptive Technologies
and
Johnson O'Connor Research Foundation

Paper presented at the Annual Meeting of
the American Educational Research Association
April, 1991

TM017555 1

BEST COPY AVAILABLE

Vocabulary research has been a pet project of the Foundation since its inception by Johnson O'Connor in the 1920s. A component of this research is used to create vocabulary instructional materials which are sold to Foundation examinees and school programs. The vocabulary department is in the process of empirically determining the difficulty values of all nontechnical words in the English language. To this end we write items for each vocabulary word and test them in public and private schools. The items are then calibrated using the Rasch model. This procedure results in a significant number of items being labelled as misfitting (not fitting the expectations of the Rasch model), and consequently being rejected from inclusion in the item bank.

In the past, items returned to the item writers as "misfitting" without any explanation regarding the cause of the problem had been a constant source of frustration. As we generate approximately 350 misfitting items per year, it was determined that creating an easy method for correcting these items would be beneficial. A "mislead analysis" technique was created in an attempt to uncover the sources of problems in items with poor fit statistics.

Fit statistics essentially reflect the mismatch of the expected response pattern of persons of known abilities with the theoretical item characteristic curve. The primary goal of the mislead analysis was to determine whether or not characteristic curves could also be constructed for the item misleads. It was

hypothesized that while most misleads would follow some sort of regular response pattern, that the offending mislead in a poorly fitting item would not match the expected response pattern.

The vocabulary items written by the Foundation follow a precise pattern where each of the misleads has a pre-defined relationship to the item, allowing separate curves to be derived for each type of mislead. The "synonym" is always the correct answer to the item and therefore the response pattern for the synonym should mimic the theoretical item characteristic curve. We assumed that the "antonym", "same situation", "similar meaning" and "sound alike" misleads would each have unique appeal to persons of differing ability levels, and therefore would each have a unique mislead characteristic curve.

The dataset used for this study contained the test results for over 3500 items each of which was administered to 400-600 persons. There were a total of approximately 23,000 persons who each took 74 items resulting in close to two-million unique person-item observations. Persons who were over 2.5 logits below the difficulty of the item were assigned to group 1, persons 1.5 to 2.5 logits below the item were assigned to group 2, persons .5 to 1.5 logits below the item were assigned to group 3, and so on until 7 ability groups were created for each item. The proportions of persons answering each item choice were then established for the 3,500 items. The proportions for each type of choice for each group were

then averaged across all 3500 items and the standard deviations computed. Figure 1 shows the means obtained for each type of choice in each of the seven ability groups. As expected, the synonym curve approximates the item characteristic curve.

The general mislead curves were then compared to the actual performance for all items which were previously labelled as "misfitting." A mislead table was constructed for each item indicating the proportion of persons who answered each item choice within a given ability group (see Table 1). The top of the table shows the distance of the persons from the items. Persons within the -.5 to .5 group are within 1/2 of a logit of the obtained item difficulty. Persons to the left are less able than the item. Persons to the right are more able. To the left of the table are the texts of each choice preceded by the type (SYN-Synonym, SIM-Similar Meaning, ANT-Antonym, SAM-Same Situation, SOU-Soundalike, BAD-choices which are blank or contain double answers). Across the bottom of the table are the total numbers of persons within each ability group. The right of the table lists the total number of persons who selected each choice. Column proportions are given only when there were at least 15 persons in the given ability group.

The mislead chart for each item also contains the item statistics provided by BIGSCALE, as well as the obtained (VSS) versus pre-estimated (LWV VSS Est)

scale score of the item (all the items are placed on the Foundation's Vocabulary Scale known as VSS).

Given the establishment of a mislead characteristic curve, one can establish the "fit" of the mislead. For our purposes we defined mislead fit by comparing the observed versus "expected" performance of a mislead for each ability group. The mislead was considered to be significantly flawed for a given ability group when the observed performance differed from the means shown in Figure 1 by more than two standard deviations. Each cell in the mislead table was evaluated in this manner, and misfitting cells marked with an asterisk for easy identification. The synonym was marked when the proportion was more than one standard deviation away from the expected value.

The item writers are now able to observe which item choices are not functioning as expected. To date, several major mislead performance patterns have been identified. Examples of each these patterns can be found in Appendices A through F.

In conclusion, what has been presented here is a simple method for helping to determine the source of misfit in vocabulary items in which the misleads are typed. Similar results were also obtained in a more recent study in which the mislead type was not known. In this case, an averaged mislead characteristic

curve was established. While the standard deviations for the values on the averaged curve were greater, the item mislead charts still provided useful information to the item writers.

Figure 1

Choice Characteristic Curves

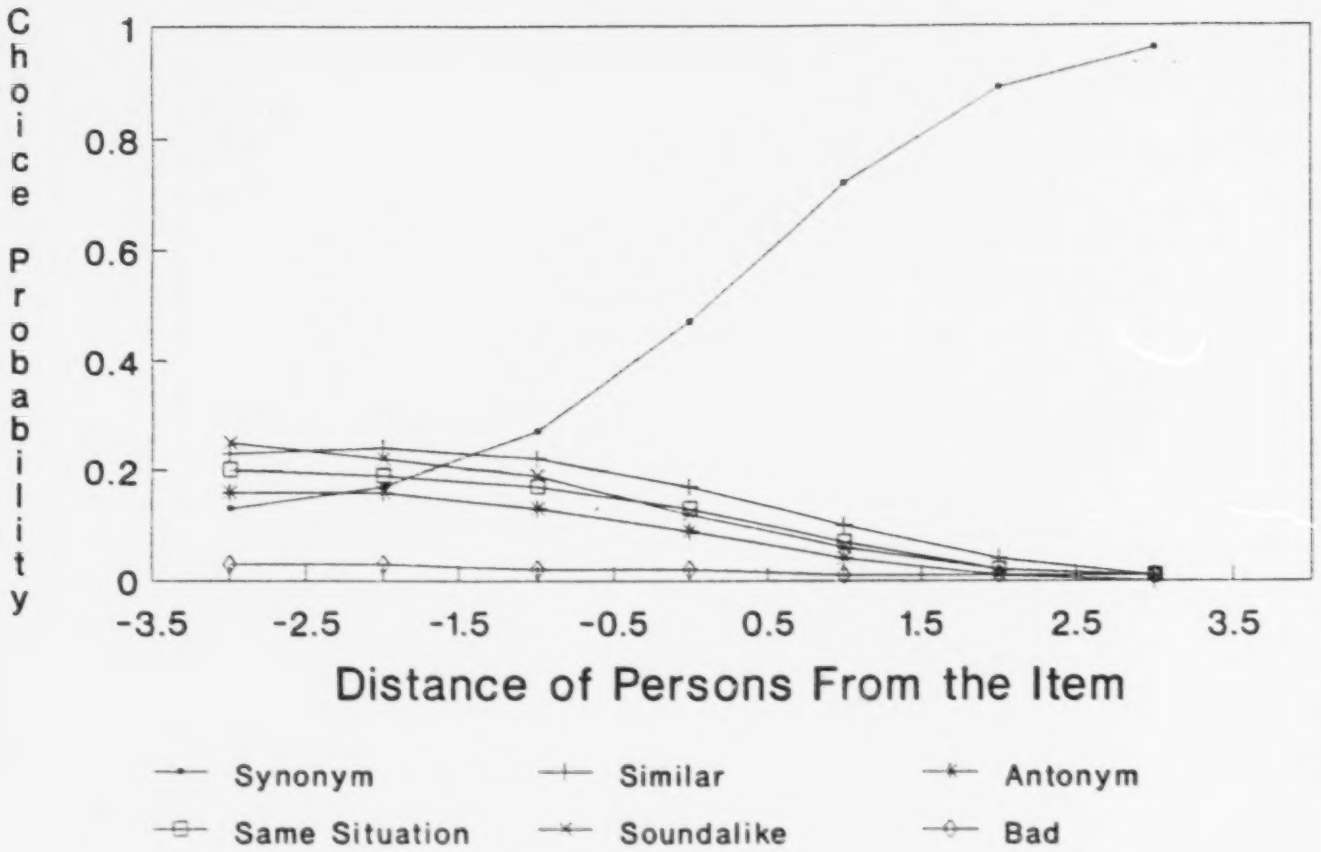

Table I

Means and Standard Deviations
for Worksample 741 Mislead Analysis

Type	-2.5 <------	-1.5 |--------	-.5 |---------	.5 |--------	1.5 |--------	2.5 |---------	>
Synonym	.13 (.04)	.17 (.05)	.27 (.06)	.47 (.06)	.72 (.06)	.89 (.05)	.96 (.06)
Similar	.23 (.12)	.24 (.13)	.22 (.13)	.17 (.11)	.10 (.07)	.04 (.04)	.01 (.02)
Antonym	.16 (.08)	.16 (.09)	.13 (.09)	.09 (.07)	.04 (.05)	.01 (.02)	.00 (.01)
Same Situation	.20 (.10)	.19 (.10)	.17 (.10)	.13 (.09)	.07 (.06)	.02 (.03)	.01 (.01)
Soundalike	.25 (.14)	.22 (.13)	.19 (.12)	.12 (.09)	.06 (.06)	.02 (.04)	.01 (.05)
Bad	.03 (.04)	.03 (.02)	.02 (.02)	.02 (.02)	.01 (.02)	.01 (.01)	.01 (.01)
N	161	805	1991	2766	2392	1282	363

N = Number of ability groups (using new items with good quality and a good sample; minimum group size = 50. Total number of persons = 22,644).

Appendix A

Inaccurate pre-estimate of the difficulty of the item resulting in the item being administered to a sample of persons with too great or too little ability relative to the item. This is observed when the ability groups are skewed to the left or the right in the table. An example of this can be found in Worksample 741, Form 3, Item 60. The word "Drive" with the synonym "Push" is a sixth grade word, but the item difficulty pre-estimate placed the word at the second grade level. The item writers probably do not need to do anything with this type of word. Instead the word should be readministered to a more appropriate sample.

741-3 Item 60. DRIVE

Obtained VSS:	127	LWV VSS Est:	[illegible]	
Measure:	-1.45	Error:	0.21	Weight: 0.23
INFIT:	0.23	Mean Square:	1.0	
OUTFIT:	2.65	Mean Square:	1.0	

Type	Text	Ability Range: <-2.5	-2.5 to -1.5	-1.5 to -.5	-.5 to .5	.5 to 1.5	1.5 to 2.5	2.5>	
SYN	push	.04*	.09*	.	.	.	.	.	24
SIM	throw	.02	.02	.	.	.	.	.	12
ANT	pull	.04	.04	.	.	.	.	.	19
SAM	lead	.11	.25	.	.	.	.	.	68
SOU	speed	.77*	.58*	.	.	.	.	.	401
BAD	(Blank/Double Answr)	.02	.02	.	.	.	.	.	10
Column Totals:		480	53	1	0	0	0	0	534

Appendix B

High ability persons selecting more than one response. This appears to occur when a mislead is either a) too close in meaning to the correct response; b) actually a second correct response (albeit not the one the item writers had intended; or c) the result of a bad key. This is probably what happened with "Mistake" (Worksample 741, Form 3, Item 48). The synonym was "fault" and the close mislead was "failure."

741-3 Item 48. MISTAKE

Obtained VSS:	55	LWV VSS Est:	9		
Measure:	-4.13	Error:	0.09	Weight:	0.02
INFIT:	4.05	Mean Square:	1.1		
OUTFIT:	3.39	Mean Square:	1.1		

Type	Text	Ability Range: -2.5	-1.5	-.5	.5	1.5	2.5	
SYN	fault	.	.19	.39*	.38*	.38*	.	.	200
SIM	lie	.	.25	.11	.09	.05	.	.	68
ANT	fact	.	.11	.07	.05	.05	.	.	38
SAM	failure	.	.22	.31	.42*	.52*	.	.	194
SOO	something lost	.	.22	.11	.05	.00	.	.	53
BAD	(Blank/Double Answr)	.	.02	.00	.00	.00	.	.	4
Column Totals:		9	64	244	219	21	0	0	557

BEST COPY AVAILABLE

Appendix C

Multiple synonyms. Any one of the misleads (not limited to the close mislead) may actually be a synonym (or near-synonym) for one of the meanings of the tested word (albeit not a true synonym of the mislead selected to be the synonym). This is probably what happened with "Cord" (Worksample 741, Form 2, Item 102). The synonym was "thick string," but the more popular soundalike was "wire," a secondary definition of cord.

741-2 Item 102. CORD

Obtained VSS:	82	LWC VSS Est:	82		
Measure:	-3.10	Error:	0.13	Weight:	0.05
INFIT:	1.73	Mean Square:	1.1		
OUTFIT:	4.05	Mean Square:	1.1		

Type	Text	Ability Range: -2.5	-1.5	-.5	.5	1.5	2.5	
SYN	thick string	.21*	.18	.19*	.21*	.	.	84
SIM	iron chain	.07	.09	.06	.00	.	.	31
ANT	fine thread	.21	.07	.02	.00	.	.	30
SAM	wire	.30	.55*	.69*	.74*	.	.	258
SOU	board	.19	.10	.03	.00	.	.	38
BAD	(Blank/Double Answr)	.03	.02	.02	.05	.	.	9
Column Totals:		73	182	176	19	0	0	450

Appendix D

Mislead not working. Though not usually a problem, it may lead to low ability persons guessing the correct answer too frequently. This occurs when low ability subjects select two or three of the choices in equal proportion, while not selecting the other choices at all, resulting in the difficulty estimate of the item being lower that it should be. The word "Beg" (Worksample 741, Form 15, Item 72) is one example of this where almost no one selected three of the misleads.

741-15 Item 72. BEG

Obtained VSS:	55	LWV VSS Est:	22		
Measure:	-4.11	Error:	0.10	Weight:	0.02
INFIT:	5.24	Mean Square:	1.2		
OUTFIT:	4.71	Mean Square:	1.2		

Ability Range:		-2.5	-1.5	-.5	.5	1.5	2.5		
Type	Text								
SYN	ask for charity	.	.	.51*	.48	.60*	.	.	231
SIM	seek	.	.	.05	.02	.04	.	.	16
ANT	donate	.	.	.05	.01	.01	.	.	10
SAM	cry out	.	.	.31	.39*	.27*	.	.	159
SOU	call to	.	.	.07	.07	.07	.	.	36
BAD	(Blank/Double Answr)	.	.	.01	.03	.02	.	.	11
Column Totals:		0	14	81	255	107	6	0	463

BEST COPY AVAILABLE

Appendix E

Synonym significantly more difficult than the test word or the misleads. This is probably what happened with the word "Hall" (Worksample 741, Form 1, Item 106). The mislead "Hall" turns out to be a second grade word and the synonym ("corridor") to be a sixth grade word. The mislead selected by the majority of persons at all ability levels was "path," which was also a second grade word.

741-1 Item 106. HALL

Obtained VSS:	101	LWV VSS Est:			
Measure:	-2.41	Error:	0.16	Weight:	0.09
INFIT:	0.75	Mean Square:	1.1		
OUTFIT:	3.77	Mean Square:	1.1		

Type	Text	-2.5	-1.5	-.5	.5	1.5	2.5		
SYN	corridor	.10	.06*	.08*	.	.	.	.	41
SIM	path	.44	.58*	.83*	.	.	.	.	279
ANT	door	.05	.02	.00	.	.	.	.	17
SAM	floor	.24	.21	.05	.	.	.	.	110
SOU	wall	.15	.12	.03	.	.	.	.	65
BAD	(Blank/Double Answr)	.02	.01	.03	.	.	.	.	8
Column Totals:		237	243	40	0	0	0	0	520